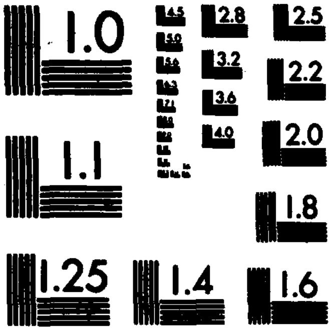
AD-A120 740 VARIABLES ON SCATTERPLOTS LOOK MORE HIGHLY CORRELATED 1/1
WHEN THE SCALES ARE INCREASED(U) HARVARD UNIV CAMBRIDGE
MASS DEPT OF STATISTICS W S CLEVELAND ET AL. JAN 82
UNCLASSIFIED PJH-12 N00014-79-C-0512 F/G 12/1 NL

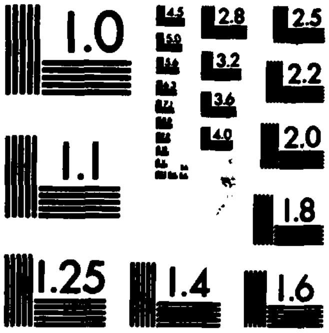
END

FILMED

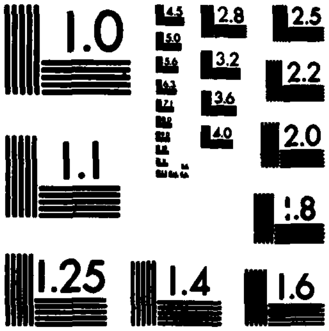
DTIC



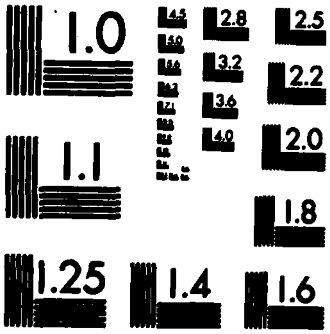
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A



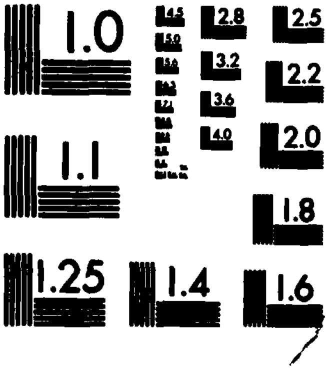
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

RESEARCH REPORT

PJH - 12

Variables on Scatterplots Look More Highly Correlated When the Scales are Increased

William S. Cleveland
Persi Diaconis
Robert McGill

January 1982

This work was facilitated in part by National Science Foundation Grant MCS-79-08685 and Office of Naval Research Contract N00014-79-C-0512.

Department of Statistics
Harvard University
Cambridge

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
PJH - 12

2. GOVT ACCESSION NO.
AD-A120740

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
Variables on Scatterplots Look More Highly Correlated When the Scales are Increased

5. TYPE OF REPORT & PERIOD COVERED
Technical Report

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
William S. Cleveland
Persi Diaconis
Robert McGill

8. CONTRACT OR GRANT NUMBER(s)
N00014-79-C-0512
NR 042-425

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Harvard University
Department of Statistics
Cambridge, MA

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS
Office of Naval Research
Code 411SP
Arlington, VA 22217

12. REPORT DATE
January 1982

13. NUMBER OF PAGES

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for Public Release; Distribution Unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Subjects were shown scatterplots and were asked to judge the amount of association between the two variables. Judged association increased when the scales on the horizontal and vertical axes were simultaneously increased so that the size of the point cloud within the frame of the plot decreased. Judges association was very different from the correlation coefficient, r, which is the most widely used measure of association.

Variables on Scatterplots Look More Highly Correlated When the Scales are Increased

William S. Cleveland
Bell Laboratories
Murray Hill, New Jersey 07974

Persi Diaconis
Stanford University,
Stanford, California 94305

Robert McGill
Bell Laboratories
Murray Hill, New Jersey 07974

ABSTRACT

Subjects were shown scatterplots and were asked to judge the amount of association between the two variables. Judged association increased when the scales on the horizontal and vertical axes were simultaneously increased so that the size of the point cloud within the frame of the plot decreased. Judged association was very different from the correlation coefficient, r, which is the most widely used measure of association.

Graphs are mainstays of the analysis and presentation of scientific data. One reason for this is that numerical summaries cannot always portray data unambiguously. For example, the most common measure of the association, or relationship, between two variables (x_i, y_i), $i = 1,...,n$, is the absolute value of the correlation coefficient, r, which measures the amount of *linear* association between two variables (1). When there is no linear association, $|r|$ is 0; when there is perfect linear association so that x_i and y_i lie along a straight line, $|r|$ is 1. However, different configurations of points can yield the same value of r, relationships can be nonlinear, and a single value of (x_i, y_i) can radically alter r (2,3). A scatterplot can depict the relationship between x_i and y_i more reliably than any single numerical measure. For this reason the scatterplot is a very commonly used tool for the investigation and presentation of the relationship between two variables. But the use of a graph opens the door for perceptual factors to enter into the analysis and interpretation of the data. While a set of data has only one numerical value for a particular measure of association such as r, the *judged* association could change according to any one of a number of "display factors" such as the size of the plotting character, the overall size of the display, the orientation of the point cloud within the frame, and the size of the point cloud within the frame. The last two factors are controlled by the scales of the vertical and horizontal axes in graphs with a fixed size plotting area.

To investigate how people judge association from scatterplots and how display factors affect their judgments, we ran three experiments. In the first experiment 74 subjects viewed 19 scatterplots, all with 0 or positive correlation coefficients. The subjects were asked to judge *linear* association on a scale from 0 to 100; 0 meant no linear association (r=0) and 100 meant perfect linear association (r=1). All subjects had some statistical training and the concept of linear association was meaningful to them. The scatterplots in Figure 1 are reductions of two of the stimuli from this experiment; the reader is invited to judge the association on these plots in order to understand the nature of the judgment task.

We varied two factors: amount of association and point-cloud size. The size of the frame was kept fixed. There were 10 levels of association; each scatterplot had a value of

$w(r) = 1 - \sqrt{1-r^2}$ equal to one of the values 0, .05, .1, .2, ..., .8. $w(r)$ is another numerical measure of linear association that goes from 0 to 1 as r goes from 0 to 1; an interpretation of $w(r)$ in terms of the geometry of the point cloud will be given later. We used $w(r)$ and not r since $w(r)$ seemed, a priori, closer to people's subjective scales than r. There were four point-cloud sizes; they are labeled 1 to 4 where size 1 is the smallest and size 4 is the largest. For point cloud size 3 there were 10 scatterplots with the 10 different values of $w(r)$, and for each of the other point cloud sizes there were 3 scatterplots with values of $w(r)$ equal to .1, .4, and .7; thus altogether there were 19 scatterplots. For both the two panels in Figure 1, $w(r) = .4$ and $r = .8$; the left panel is point-cloud size 2 and the right panel is point-cloud size 4.

Each scatterplot had 200 points and a square frame with sides equal to 17.3 cm. In all cases the center of gravity of the point cloud was at the center of the frame. The values portrayed on the horizontal axis of the k-th scatterplot, $x_i(k)$, for $i = 1,...,200$, and the values portrayed on the vertical axis, $y_i(k)$, for $i = 1,...,200$, formed a bivariate super-normal point cloud (4) which insured highly regular behavior: a linear relationship, no peculiar points, and an elliptical appearance. The major axis of each point cloud was the line $y = x$ and the minor axis was the line $y = -x$.

The minimum value portrayed on the two axes of all plots was 0 data units and the maximum value was 5.6, 7, 10, or 14 data units. Since the length of each axis was 17.3 cm, the four scale values were .32, .40, .58, and .81 data units/cm. The effect of decreasing the scale was to increase the size of the point cloud within the frame.

There were 4 orders of presentation of the 19 scatterplots with approximately 1/4 of the subjects judging each order. Two of the orders were random and the other two were the reverses of these.

Subjects judged the scatterplots in stapled booklets with 8-1/2"X11" pages. First there were written instructions and sample scatterplots, then four trial plots that subjects judged, and

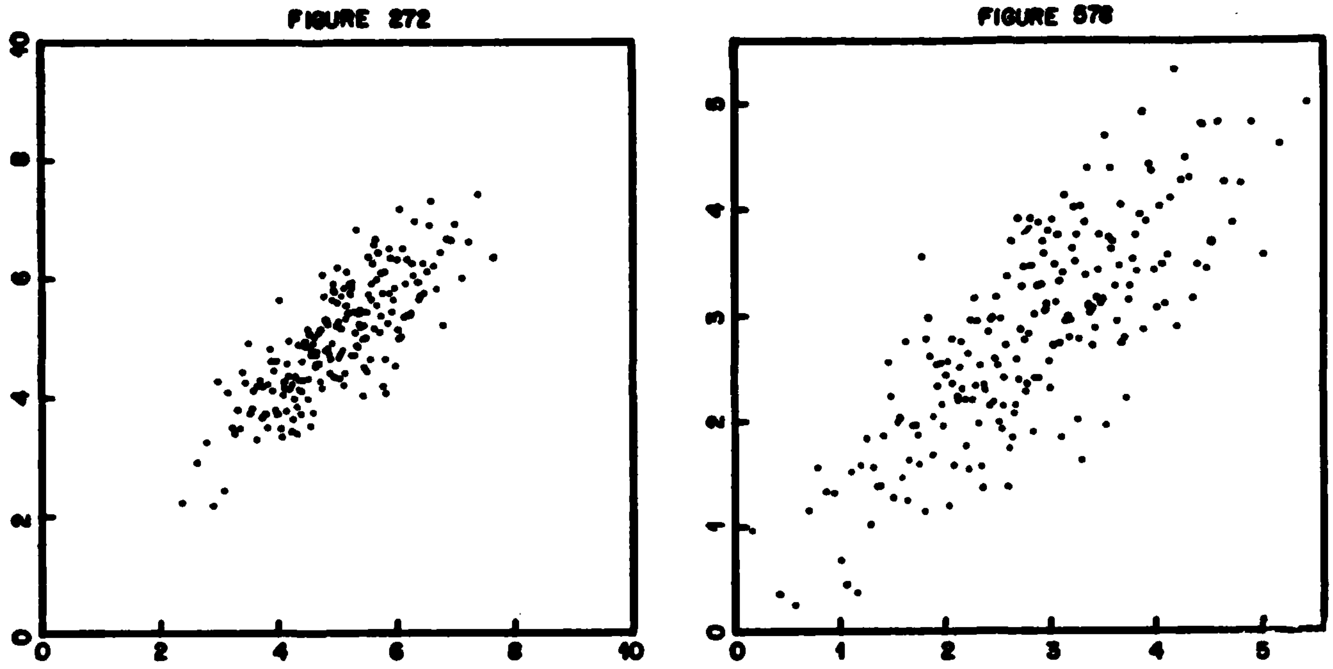


Figure 1. Reductions of two scatterplots used in the three types of experiments. The left panel is point-cloud size 2 and the right panel is point-cloud size 4.

finally the 19 experimental plots, each on a separate page. Subjects were asked to give their own subjective assessment of the amount of linear association, rather than to judge the correlaaion coefficient, and were asked not to look back or change old answers. It was suggested that they work reasonably quickly and that most people could comfortably make a single judgment within 15 seconds. Subjects, all of whom had a basic knowledge of statistics, fell into three categories: students taking university courses in statistics, university faculty members in statistics and mathematics, and statisticians practicing statistics in government and industry.

Our data analyses made extensive use of 10% trimmed means (5), which are defined in the following way: Order the observations from smallest to largest; drop the largest 10% of the observations and the smallest 10%; take the arithmetic average of the remaining values. 10% trimmed means are robust estimates (6,7) since they are not distorted by a small fraction of outliers, and they are a compromise between arithmetic means, which are 0% trimmed means, and medians, which are trimmed means close to the 50% level. The standard errors of 10% trimmed means can be computed from a formula given in (8).

Judged association for each of the 19 scatterplots was summarized by 10% trimmed means of the subjects' guesses, which were on a scale of 0 to 100, divided by 100. These values are plotted in Figure 2 against the actual values of r for the 19 scatterplots; also portrayed are the standard errors of the trimmed means. The two curves are $w(r)$ and $g(r) = (1-r)/(1+r)$. $g(r)$ is another measure of linear association that goes from 0 to 1 as r goes from 0 to 1; an interpretation of $g(r)$ in terms of the geometry of the point cloud will be given later.

Figure 2 shows that judged association is quite different from the standard numerical measure, r, since the 10% trimmed means lie well below the line $y = x$. This result has been found in two other experiments (9,10) in which subjects were asked to guess the correlation coefficient from scatterplots and in experiments in which the amount of association was judged on the basis of other kinds of stimuli (11). Interestingly, these results also correspond to a

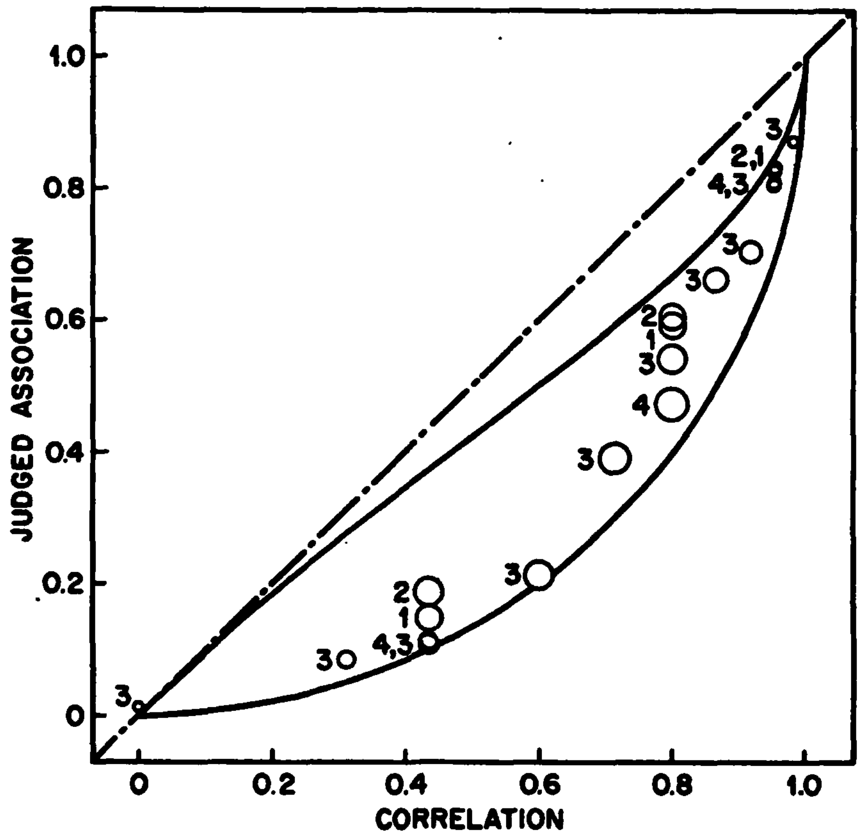


Figure 2. The 10% trimmed means across subjects of judged association divided by 100 for 19 scatterplots are plotted (by the circle centers) against the values of r, the correlation coefficient, of the scatterplots. The circle radii portray the standard errors of the trimmed means. Thus the circle areas are proportional to the estimated variances of the trimmed means. The numbers to the left of the circles indicate the point-cloud sizes. When two numbers are shown, separated by a comma, two circles are nearly coincident and the first number refers to the circle with the smaller trimmed mean. The upper solid curve on the plot is $g(r)$ and the lower solid curve is $w(r)$. The dashed line is the line $y = x$. The information on the plot leads to two conclusions: judged association tends to increase as the point-cloud size decreases due to increasing scale; judged association is very different from the standard numerical measure of association, the correlation coefficient.

statement of Wilk (12): "... it is felt by some [applied statisticians] that values of $|r|$ below .5 are quite 'small', while r is 'large' only when $|r|$ is above .8, and r is 'really large' (close to a linear dependence of the variables) only when $|r|$ is above .95." Wilk argues that $w(r)$ is a more sensible numerical measure of association than r; Figure 2 shows that $w(r)$ does come closer to describing the perceived association for our subjects than does r.

Figure 2 also shows that the tendency is for judged association to increase as the point cloud size decreases due to the increase in the scales; the effect is most pronounced when $w(r) = .4$. In all cases the perceived associations for sizes 1 and 2 are greater than those for sizes 3 and 4. The effect, however, does not appear to extend beyond size 2; for all three values of $w(r)$, the trimmed mean for point-cloud size 2 is either very close to that of size 1 or somewhat greater. And sizes 3 and 4 differ from one another by a nontrivial amount only for $w(r) = .4 (r = .8)$.

To investigate the statistical significance of the effect of changing scale we performed the following operations: For each subject and each level of $w(r)$ in which scale was varied ($w(r) = .1, .4, .7$) we subtracted the subject's estimate for the largest point-cloud size, 4, from each of the estimates for the other three sizes, which for each subject yielded 3 differences for each of the 3 levels of $w(r)$; then we computed 10% trimmed means and their standard errors across the subjects. Each trimmed mean divided by its standard error has, approximately, a t-distribution with 57 degrees of freedom (8); this distributional result can be used to test the significance of the difference of the point-cloud size 4 response from the responses to the other sizes. For $w(r) = .1$ the size 4 response is significantly different (at the .01 level) only from the level 2 response; for $w(r) = .4$ the size 4 response is significantly different from all three of the other responses; for $w(r) = .7$ the size 4 response is significantly different only from the size 1 response.

We ran a second experiment to check, under different conditions, this effect of scale. 109 subjects in 3 groups of 27, 36, and 46 people were shown, alternately, the two scatterplots in Figure 1 by an overhead transparency projected onto a screen in the front of a room. Subjects were asked to assess the association of each plot on a scale of 0 to 100. The 10% trimmed means of

$$\frac{\text{(judgment for point-cloud size 2)} - \text{(judgment for point-cloud size 4)}}{100}$$

across subjects is .068 with a standard error of .011. The 10% trimmed mean of the corresponding values for the subjects in the first experiment is .125 with a standard error of .018.

We ran a third experiment to further investigate the results. Thirty-two subjects in a single group were shown the scatterplots in Figure 1 in the same manner as the subjects in the second experiment. But in this case subjects were told that the correlation coefficients of the two scatterplots were the same and were asked to indicate whether one of the two "looked" more highly correlated than the other and if so, which one. 66% indicated that the size 2 scatterplot looked more correlated, 13% indicated the size 4 scatterplot, and 22% said they looked the same. This has the same pattern as in the first experiment, where the corresponding percents are 81%, 18%, and 15%, and in the second experiment, where the corresponding percents are 59%, 11%, and 30%.

Thus the second and third experiments strongly corroborated the conclusion of the first experiment: increasing the scales on the horizontal and vertical axes of a scatterplot so as to decrease the point-cloud size, increases the judged association.

Knowing what perceptual strategies people employ in judging association from scatterplots might not only provide an explanation of the effect of scale in our three experiments, but might also enable more effective design of scatterplots. The point clouds on the scatterplots in our experiments have an elliptical look to them because the bivariate normal distribution, from which we can think of the points as arising, has a density with elliptical contours. Two of the features of the point clouds are the ratio of the lengths of the minor and major axes and the area. Subjects might be using either to judge association.

The ratio of the minor axis to the major axis of a contour of the associated bivariate normal distribution is $(1-r)/(1+r)$, since the standard deviations of $x_i(k)$ and $y_i(k)$ are equal and since the scales on the horizontal and vertical axes of each scatterplot are the same. If subjects were judging association by judging the ratio of the axes of the point-cloud, then the judged scale would be $g(r)$, which, as described earlier, is shown in Figure 2.

The area of an elliptical contour of the associated bivariate normal distribution divided by the area of a rectangle with sides parallel to the horizontal and vertical axes of the plot, is equal to $\sqrt{1-r^2}$. If subjects were judging association by judging the areas of the point clouds relative to a circumscribed rectangle, the judged scale would be $w(r)$, which, as described earlier, is shown in Figure 2.

Neither of the curves $w(r)$ and $g(r)$ appear to describe the judged association. It could be, however, that one of the two geometrical tasks — judging axis ratios or judging areas — is being carried out, but that there are biases in the judgments that alter the perceived association. For example, it is known that judgments of area and length tend to be proportional not to the physical quantity, but rather to the physical quantity to a power less than 1 (13). We have begun a series of new experiments to attempt to better understand the perceptual mechanism that people use in judging association.

References and Notes

(1) G. W. Snedecor and W. G. Cochran, *Statistical Methods* (The Iowa State University Press, Ames, Iowa, 1967).

(2) S. J. Devlin, R. Gnanadesikan, and J. R. Kettenring, *Biometrika*, 62, 531 (1975).

(3) F. J. Anscombe, *The American Statistician*, 27, 17 (1973).

(4) Let q_i for $i = 1,...,200$ be equally spaced quantiles of the normal distribution so that $\Phi(q_i) = (i-.5)/200$. Let u_i be the q_i divided by their standard error. The values

portrayed on the horizontal axis of the k-th scatterplot are $x_i(k) = \alpha(k) + \beta u_i$ for $i = 1,\ldots,200$. Let $v_i(k)$ be a random permutation of the u_i; linearly regress $v_i(k)$ on u_i and let $w_i(k)$ be the residuals divided by their standard error. Let r_k be the desired correlation coefficient of the k-th scatterplot. The values portrayed on the vertical axis of the plot are $y_i(k) = \alpha(k) + \beta[r(k)u_i + (1-(r(k))^2)^{-.5} w_i(k)]$. Both $x_i(k)$ and $y_i(k)$ have standard deviation β and their correlation is $r(k)$. A different permutation is used for each scatterplot. The $\alpha(k)$ are chosen so that the center of gravity of the point cloud is at the center of the plotting frame. β has a value that places the extremes of the point cloud for the smallest scale value just inside the plotting frame.

(5) F. Mosteller and J. W. Tukey, *Data Analysis and Regression* (Addison-Wesley, Reading, Mass., 1977).

(6) P. J. Huber, *Robust Statistics* (Wiley, New York, 1980).

(7) F. R. Hampel, *Journal of the American Statistical Association*, 69, 383 (1974).

(8) J. W. Tukey and D. H. McLaughlin, *Sankhya*, Ser. A, 25, 331 (1963).

(9) R. F. Strahan and C. J. Hansen, *Applied Psychological Measurement*, 2, 543 (1978).

(10) P. Bobko and R. Karren, *Personnel Psychology*, 32, 313 (1979).

(11) D. L. Jennings, T. Amabile, and L. Ross, *Judgment Under Uncertainty: Heuristics and Biases*, edited by A. Tversky and D. Kahneman (Cambridge University Press, New York, 1980).

(12) M. B. Wilk, *Bell Laboratories Technical Memorandum* (Bell Laboratories, Murray Hill, N.J., 1966).

(13) S. S. Stevens, *Psychophysics — Introduction to its Perceptual, Neural, and Social Prospects* (Wiley, New York, 1975).

(14) We are greatly indebted to Ram Gnanadesikan, Colin Mallows, Saul Sternberg, Walter Tapp, and Paul Tukey for many helpful comments on an earlier manuscript.